ON BROWSING: THE USE OF SEARCH THEORY IN THE SEARCH FOR INFORMATION

by

Philip M. Morse

Technical Report No. 50

OPERATIONS RESEARCH CENTER

Reproduced by the
CLEARINGHOUSE
for Federal Scientific & Technical
Information Springfield Va. 22151

MASSACHUSETTS INSTITUTE
OF
TECHNOLOGY

February 1970

ON BROWSING: THE USE OF SEARCH THEORY IN THE SEARCH FOR INFORMATION

by

PHILIP M. MORSE

Technical Report No. 50

Work Performed Under

Contract No. DA-31-124-ARO-D-209
U. S. Army Research Office (Durham)

Dept. of the Army Project No. 20011501B704
DSR 75217

Operations Research Center
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Cambridge, Massachusetts 02139

February 1970

FOREWORD

The Operations Research Center at the Massachusetts Institute of Technology is an interdepartmental activity devoted to graduate education and research in the field of operations research. The work of the Center is supported, in part, by government contracts and industrial grants-in-aid. Expenditures associated with the work reported herein were supported by the U. S. Army Research Office (Durham) under Contract No. DA-31-124-ARO-D-209.

John D. C. Little
Director

Browsing may be defined as a search, hopefully serendipitous. In connection with a library, one may browse through the display of recent books to see what is new, or through a portion of the library shelves in the hope of finding a text which might contribute the fact or idea needed in some intellectual effort. One might scan quickly through the fiction collection to see whether some title might strike one's fancy or, more rarely, might thumb through the card catalogue to see whether some known author has written a book one has not yet read. In each case the browser is not certain he will find anything of use to him but he has hopes, and past experience supports that hope. Browsing is prevalent in most libraries. In fact it can be argued that browsing is one of the most frequent ways in which the library user finds the books he borrows. To analyze browsing probabilistically, to see whether browser or librarian can improve its efficiency, one might try applying the theory of search.

Search Theory

Search theory was developed in World War II in connection with antisubmarine warfare[1]. Probability theory and geometry suggested, and experimental observation verified, that there was a fairly simple relationship[2],between the chance of success in spotting a submarine in a given area of the ocean, and the degree of effort spent by a patrol aircraft, for example, in searching the area. If the submarine is somewhere in area A then the probability of success P_s in spotting the submarine is

$$P_s = 1 - e^{-\phi} \qquad (1)$$

where ϕ, the search coverage, equals $\rho T/A$, the search rate of the plane in square miles per hour, multiplied by T the hours spent in the area and divided by the number of square miles in area A ($e = 2.718$ is the base of natural logarithms). The search rate ρ depends on the altitude of flight of the plane, its speed and on the search method (radar or visual) and equipment; it has to be measured for each plane and equipment.

Figure 1 is a plot of P_s versus ϕ. Note that even though the area is "covered" (i.e., $\phi = 1$) it still is not certain (i.e., P_s is not unity) that the submarine is spotted, even though it is there and on the surface. Errors in navigation will leave some areas uncovered while other areas are "oversearched"; operators and equipment are fallible. Poor planning and maintenance often lowered the chance of success below that given in Eq. (1); very seldom was it bettered. Note also that, in general, the coverage is proportional to the time spent. It usually turned out that using a faster plane, to search the same area in shorter time, simply increased the number of times the target was overlooked. Particularly in the case of visual search, experiments made during the war by Selig Hecht (unpublished) showed that "haste makes waste". Coverage ϕ in general was proportional to the time spent per unit area of scan; it didn't matter much whether this time was spent by covering some subarea thoroughly or else by scanning cursorily over the whole area.

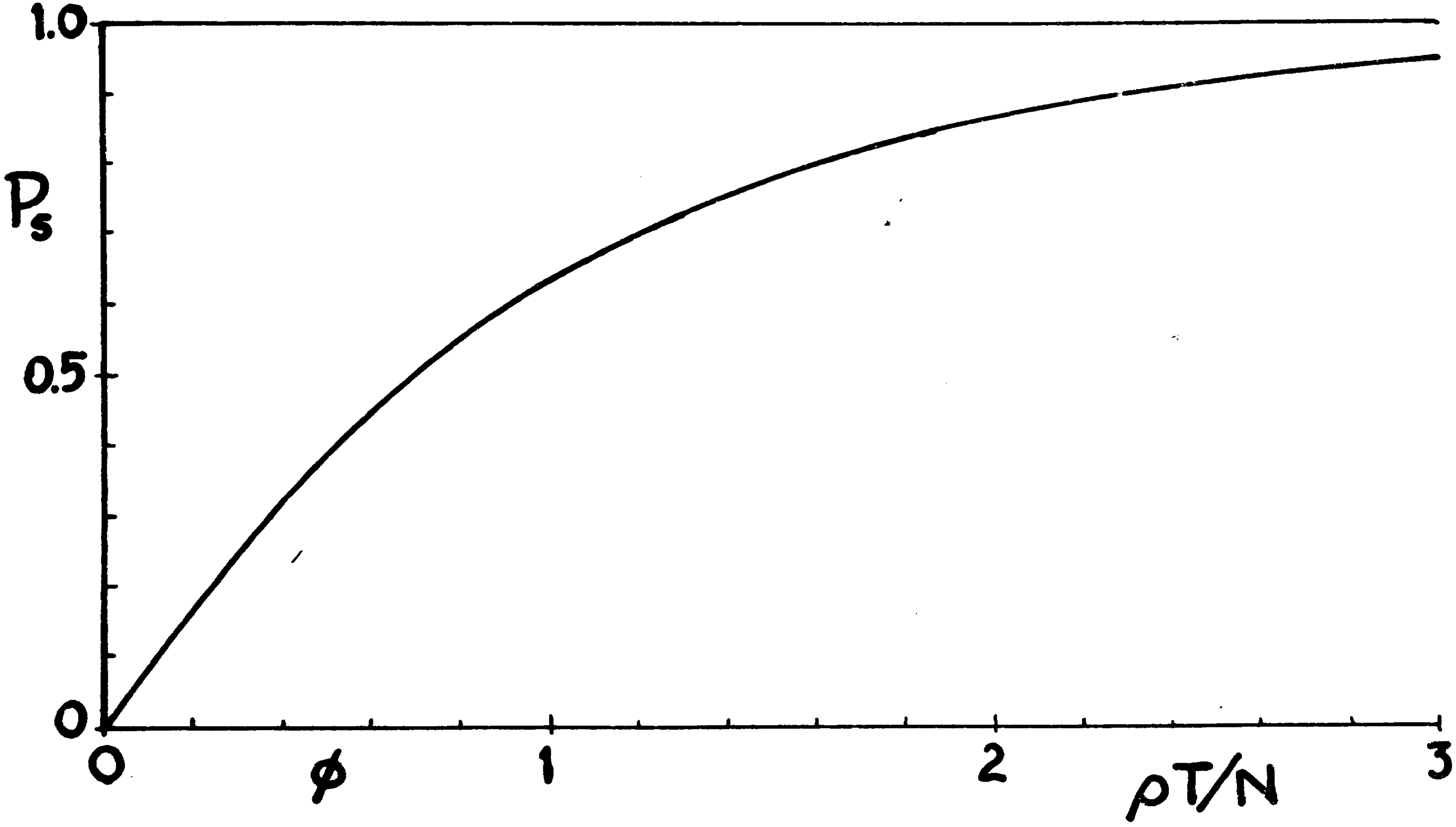

Fig.1 Chance of spotting one book of present interest, placed at random in N books, as function of search parameter ρT/N.

If there are two areas of the ocean, A_1 and A_2, if the probability that the submarine is in A_1 is p_1 and the probability that it is in A_2 is p_2,then the chance the submarine is spotted is

$$P_s = p_1(1 - e^{-\phi_1}) + p_2(1 - e^{-\phi_2}) \qquad (2)$$

The search coverage ϕ_1 of area A_1 is $\phi_1 = \rho T_1/A_1$, with T_1 the time spent in A_1; similarly for $\phi_2 = \rho T_2/A_2$, with T_2 the time spent in A_2. Formulas were developed[2] determining optimal allocation of search time between the two areas, in order to maximize P_s.

Search and Browsing

Let us now apply search theory to the "operation" of browsing, of scanning the books on a set of shelves in the library. Suppose the shelves contain N books. The chance that the browser will spot a particular book, placed at random among the N books, is

$$P_s = 1 - e^{-\phi} \quad ; \quad \phi = \rho T/N \qquad (3)$$

where T is the time spent and ρ is a constant that might be called the <u>search rate of the browser</u>. Its value varies from person to person and also depends on the accessibility and illumination of the shelves. Its value for a particular browser and set of shelves can be determined by running a series of trials (20 or more), each run for the same time T, each with a different target book, placed at random in the collection, to see what fraction P_s of the trials end in finding the book within time T (N should be at least 1000 and T should be chosen so P_s is between 1/3 and 2/3 for best accuracy). Knowing N and T and estimating P_s from the trials, ρ can be obtained. For the purposes of this analysis an estimate within a factor of two is sufficient. Measurements made by the writer indicate that, for him, under good lighting conditions, ρ is somewhere between 100 and 200 volumes per minute.

To apply this formula to browsing we have to reach some conclusions regarding the book (or books) searched for. In most cases the browser does not know himself which book he will pick out, nor indeed whether he will find any book he wants just then, even if he spends all day at it. Nevertheless each regular user of a library has some inkling of which portion of the library is more likely to yield books of immediate interest to him. If put to it, by using methods developed by decision theorists[3], he could estimate a priori an expected number E of books, of interest to him at the moment, that might be present in a specific section of N books, though he does not know where in the section the books might be(nor, ahead of time, just what book it might be). For the purpose of this paper it is sufficient if he can estimate E to within a factor of 2 or 3. Habitual browsers in a library do this intuitively when selecting which section of the library they will browse in during a particular stay. They go to that section of the library which they estimate has the greatest likelihood of having a book they might want just then to read. Of course immediate interests change; a particular browser may have a completely different set of values of the E's next time he visits the library.

Thus our theory indicates that the browser, during a given visit, may divide his search among M different sections of the library (perhaps distinguished by general subject matter or location), spending time T_1 in the first section, which contains N_1 books, and so on for the M sections. If he does this, the expected number of books he will find of immediate interest to him is the sum

$$S = S_1 + S_2 + \cdot \cdot \cdot \cdot + S_M \qquad \text{where}$$
$$S_m = E_m(1 - e^{-\phi_m}) \quad ; \quad \phi_m = \rho T_m/N_m \qquad (4)$$

As mentioned earlier, E_m is his <u>a priori</u> estimate of the number of books of immediate interest to him which might be in section m and ρ is his search rate (ρ may vary from section to section, but this complication is not usually worth adding). Of course, in any particular browse, he may not find <u>any</u> books of interest in section m, or he may find $3S_m$; search theory indicates that S_m is his best <u>a priori</u> estimate of what would be the result of his spending time T_m in the m'th section. We might emphasize the probabilistic nature of S_m by calling it the <u>expected</u> <u>success</u> of his proposed expenditure of time in scanning the m'th section. The total time spent browsing, during that particular visit, is of course the sum $T = T_1 + T_2 + \cdots + T_M$.

<u>The Browser's Problem</u>.

On the basis of his estimates of E_m the browser has the problem of distributing the total time T he wishes to spend, in such a manner as to make the total expected success S as large as possible. That this is a meaningful problem is due to the fact that search is subject to the law of diminishing returns. Figure 1 shows that doubling the time spent scanning a given section does not double the expected success. In fact if enough time has already been spent so that $\phi = \rho T/N$ is larger than 2, increasing ϕ to 4 by spending another equal amount of time in the same section can only increase S by about another ten percent; certainly it would be better to spend this additional time in scanning another section, of equal promise, as yet unscanned.

This can be made precise by asking what division of total time T should be made between two sections of equal number of volumes N each and with equal estimated numbers E of books of immediate interest to the browser. The expected value of S, if the visitor spends time t in one section and time T - t in the other, is

$$S = E(1 - e^{-\rho t/N}) + E(1 - e^{-\rho(T-t)/N})$$

which is plotted in Fig. 2 for $\rho T/N = 2$. It is obvious that the maximum is reached when time T is divided equally between the two sections of equal promise E, though the flatness of the maximum indicates that it is not very important that the equality be precise. The symmetry of the figure indicates that as long as N and E are equal in the two sections, the time should be equally divided, no matter how large or small is T, the total time to be spent browsing. The statement can be extended: if there are M sections, all equal in regard to E and to N, then the total browsing time T should be divided equally among all M sections, spending time T/M on each, to achieve maximum expected success. Indeed, if one has no idea what might be useful, so that the density of expected books of interest, E/N, is, a priori, the same for every section in the

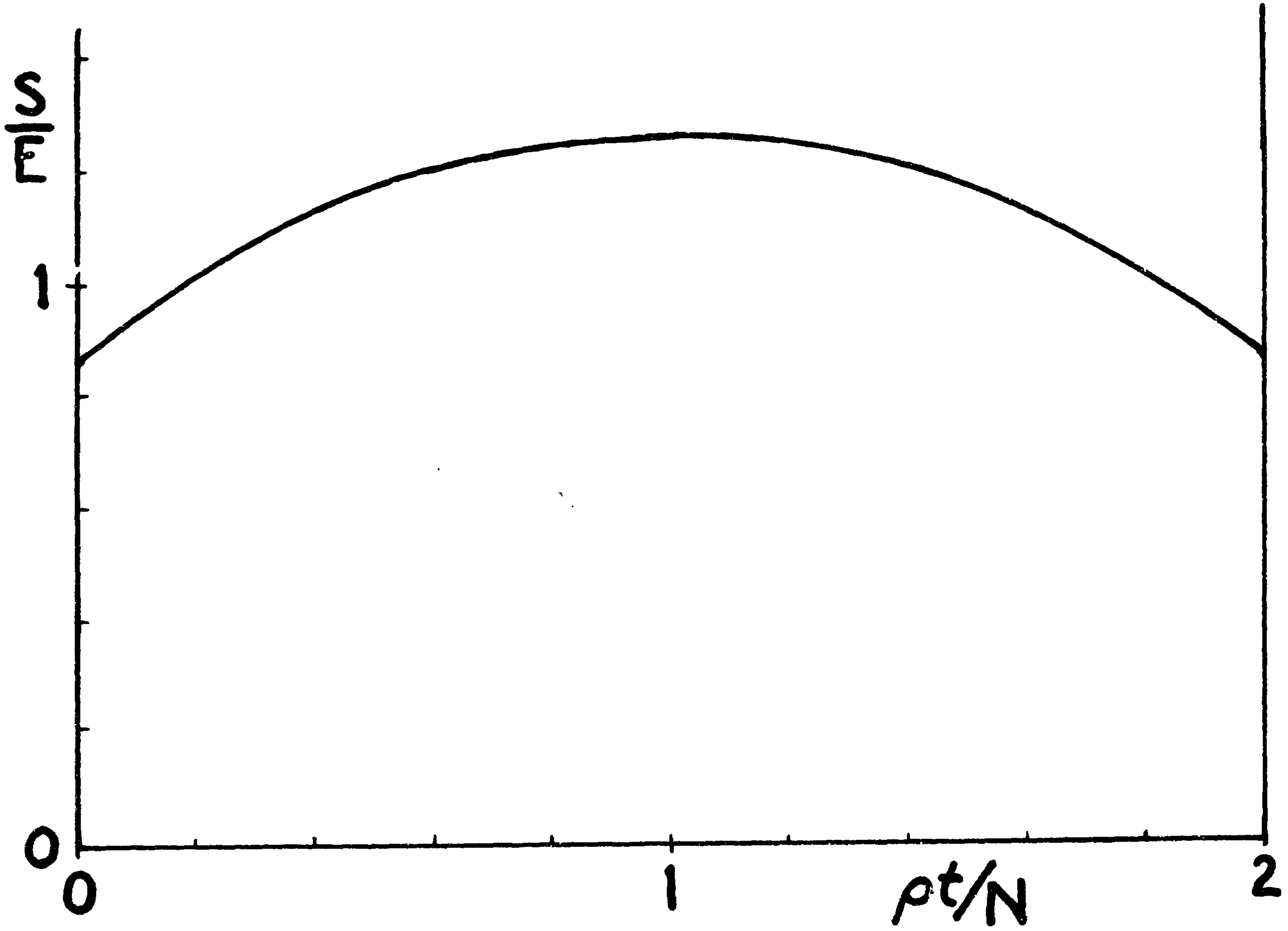

Fig.2. Expected success as function
of relative allocation of search between
two sections of equal promise.

library, then it is more productive to scan, rapidly and more or less uniformly, all the sections rather than concentrate on one portion[4].

We note that it is the <u>estimated</u> <u>density</u> of books of possible immediate interest, $V = E/N$, which is the criterion. If this density is uniform in a section or sections of the library, if the expected book or books of possible immediate interest are equally likely to be <u>anywhere</u> in the section or sections, then the browser should spend roughly equal time scanning each portion of them, even if this means only a quick scan along each shelf. Probabilistically we can say that each book in the section or sections has an equal <u>a</u> <u>priori</u> <u>interest</u> <u>potential</u> $V = E/N$ for this browser for this visit, and thus deserves an equal portion of the scan (until, of course, a satisfactory number of books of interest have actually been found).

A more difficult problem arises when the estimated density, or interest potential V varies from section to section of the library; what then should be the allocation of time spent in browsing? The derivation of the formula is given in the reference[2]; here we need only translate the result into terms appropriate for book browsing. It will be more understandable if we start by applying it to a specific case.

Suppose there are four sections which have promise for the potential browser this visit. The first section has $N_1 = 1000$ volumes and the <u>a priori</u> estimate is that it might contain $E_1 = 3$ books of immediate interest. Since he has no idea where these 3 books might be in the section, the prospective browser must assume (until he finds otherwise) that each book in this section has an interest potential $V_1 = 3/1000 = 0.003$. The values of N, E and V for each of the four sections are

TABLE I

Sect.	N	E	V	lnV	lnV + 7.3	lnV + 8.22
1	1000	3	0.003	-5.8	1.5	2.42
2	5000	2	0.0004	-7.8	-	0.42
3	5000	1	0.0002	-8.5	-	-
4	9000	1	0.00011	-9.1	-	-

$\ln(0.00069) = -7.3$; $\ln(0.00027) = -8.2$

given in Table I. We first look up the natural logarithms of the interest potentials V of the books in each section; this is given in the fifth column of Table I. We will assume, for the purpose of the example, that the browser's search rate ρ is 150 books per minute.

The situation is more understandable as shown in Fig.3, where we have plotted four rectangles, of width proportional to the respective number of books in each section, and of height equal to the corresponding lnV. Now suppose the prospective browser has only $T = 10$ minutes to spend browsing; how should he divide his time among the sections? We obtain the solution by drawing a horizontal line, at the level marked $\ln\lambda_{10}$, such that the area between this line and the top, heavy line of the plot is just $\rho T = 150 \times 10 = 1500$. This area is reached for $\ln\lambda_{10} = -7.3$, when the area, cross-hatched in the figure, $(\ln V_1 - \ln\lambda_{10})N_1 = (-5.8 + 7.3) \times 1000$ is just equal to 1500. In this case only section 1 is involved; the prospective browser should spend <u>all</u> his 10 minutes in section 1; he maximizes his

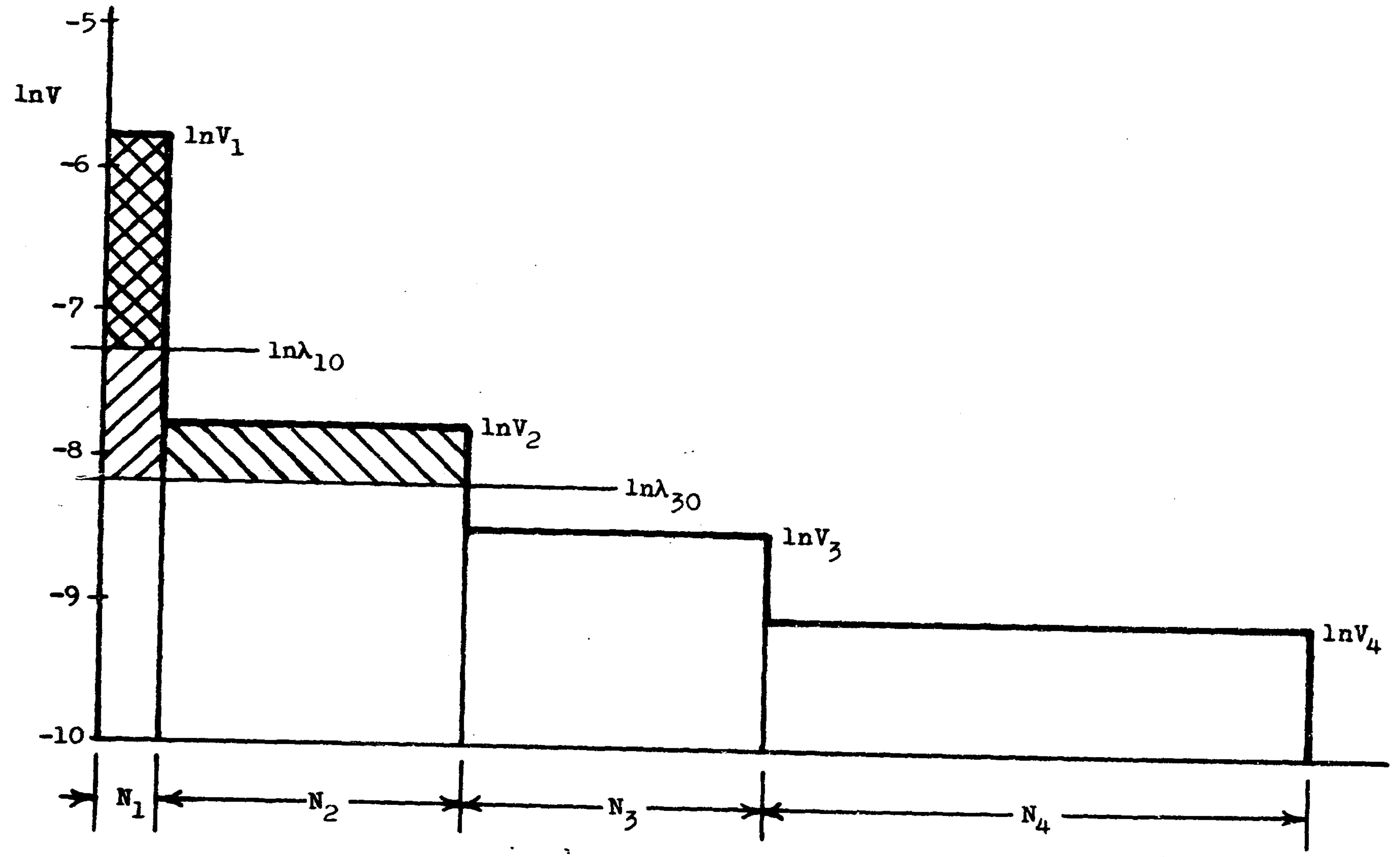

Fig.3. Graphical illustration of optimal allocation of search effort.

expected success by ignor:ng the sections with lower interest potential V. The expected success,

$$S = E_1 - \lambda_{10}N_1 = 3 - 0.00069\times1000 = 2.3 \text{ books}$$

is greater than could be attained by devoting any part of the ten minutes to any of the other sections.

Next suppose the browser has 30 minutes at his disposal. In that case we place $\ln\lambda_{30}$ so that the enlarged area (that shaded plus that cross-hatched) just equals $\rho T = 150\times30 = 4500$. This happens to come at $\ln\lambda_{30} = -8.2$, which has some area in section 1 and some in section 2, but none in sections 3 or 4. The relative times are to be divided in proportion to the areas

$$(N_m/\rho)(\ln V_m - \ln\lambda_{30})$$

time in 1 $= \frac{1000}{150}(-5.8+8.22) = 16$ minutes, evenly spread over 1

time in 2 $= \frac{5000}{150}(-7.8+8.22) = 14$ minutes, evenly spread over 2

Thus he should spend less time in section 2, even though there are 5 times as many books there, and he should still ignore sections 3 and 4. Only when he has more than about 41 minutes should he start glancing at section 3 and only if he has more than 85 minutes to browse should he bother with section 4. With the 30 minute limit and the 16 - 14 division, his expected success is $S = E_1 - \lambda_{30}N_1 + E_2 - \lambda_{30}N_2 = 3 - 0.27 + 2 - 1.35 = 3.4$ books which is the maximum he can expect to find (though he may be lucky some times, of course).

To generalize the procedure, we plot the situation as in Fig. 3, with the m'th section represented by a column of height $\ln V_m = \ln(E_m/N_m)$ and with width equal to N_m. We then find the horizontal line, at level $\ln\lambda_T$, for which the area between it and the top line of the plot is just equal to ρT, with T equal to the time available for browsing. The time T_m to be spent in section m, and the total expected success S are then given by the equations

$$T_m = \begin{cases} (N_m/\rho)(\ln V_m - \ln\lambda_T) & (\text{if } V_m > \lambda_T) \\ 0 & (\text{if } V_m < \lambda_T) \end{cases} \qquad (5)$$

$$S = E_1 - \lambda_T N_1 + E_2 - \lambda_T N_2 + \cdot \cdot \cdot$$

where the sum for S includes only those sections for which $V_m = E_m/N_m$ is greater than λ_T.

Of course it would be foolish for the expectant browser to go through such an analysis in detail before he starts browsing (he would use up all his time just figuring out what to do!). However the essential point of the exercise is that wherever the interest potential V = E/N is uniformly spread then the search should be uniformly spread; and wherever the interest potential is higher than in other sections there the search effort should be *strongly concentrated*, *even* to the extent of *ignoring entirely* sections of lower potential. Thus search allocation should be a non-linear function of interest potential. Of course if it is estimated that one portion of a section has higher interest potential than the rest, then this portion should be considered a separate section (for this browser) to be searched much more carefully than the rest.

The primary purpose of this analysis has been to make us familiar with the methods and concepts of search theory, as applied to libraries; now we can go on to discuss the more important problem, of what the librarian can do to improve matters for all browsers.

The Librarian's Problem

The problem is relatively simple for each individual library user. Though his desires may change from visit to visit, he needs only to estimate the interest potential of books in various parts of the library, in accord with his immediate interests, and then to allocate his search efforts as has been outlined, concentrating strongly on the highest potential areas. It is quite otherwise for the librarian, for the interests of different browsers differ widely; indeed the interests of the same browser vary widely from visit to visit. Is there anything the librarian can do to improve the success of all browsers, or at least to improve the success of the average browser?

One thing is apparent immediately; the librarian should arrange his collections so as to be obviously differentiable, in interest potential, to the majority of library users. The worst imaginable library, for a browser, would be one in which he could not differentiate at all between the interest potentials of different sections, where he would have to treat all shelves as being equally likely (or, rather, equally unlikely) to produce what he might want. That library which makes it possible for the average browser to pick quickly a relatively few, relatively small sections of high interest potential for his present desires, so he can ignore the rest, is the library which is most efficient for the browser to use. The subdivision is not easy; too fine a division makes it necessary for the browser to search too many sections in order to cover his range of interest; he should not have to cover more than about three sections per trip.

¶ Parenthetically, this is the reason why card catalogues are absurdly ineffective for browsing. Aside from the very small search rate, every drawer is more or less equally sparse in interest potential; very few interest spans go according to the alphabet, even in the subject catalogue. It is important for the designers of computerized catalogues to realize that such a catalogue also will be spurned by the browser (and thus will have its usefulness seriously impaired) if it does not provide for *quick* and *simple* means of assembling sub-catalogues of high interest potential, *no matter what the interest span* of the potential browser may be. If the computer can assemble, in a minute or so, a sub-catalogue of a few thousand items, all of high interest potential, of combinations of such disparate subjects as entire functions, decision theory, ideas of probability in Hellenic mathematics and/or data on book-use in college libraries, for example, with the browser then able to flick through the collection in five or ten minutes; only then will the computerized catalogue *begin* to replace the simple roaming through the stacks, which has always been (and *may* always be) the usual way of finding what book one wants.

But to return to the librarian's present problem. He will (and does) help the browser immensely by arranging the books on the shelves, not alphabetically, or at random, but by "subject class", so if the user knows his Dewey or LC code he can quickly pick out the high-interest-potential regions for his present predelictions. The trouble comes when the collection becomes too large for all of it to be easily available to all, when even one class becomes so large it cannot be scanned efficiently in a fraction of an hour.

By this time, of course, even in one subject class, the high-interest-potential items have been diluted with a lot of old and/or highly specialized books, of interest to very few library users, which lower the interest potential of each section, for nearly all users. This may be of little moment in some research libraries, where browsing is seldom practiced, but in most libraries this would mean that browsing is no longer efficient and hence is frustrating. By this time also, it usually happens that the collection has got so large that it cannot all be kept in one place. The question therefore is, how to subdivide the subject classes so that one portion regains its original browsing effectiveness, without at the same time destroying other utilities.

One possibility is to subdivide by subject matter, to establish instead of a science library, for example, a physics library, a mathematics library, etc. But this solution further reduces the browser's chance of success. For if the subject groupings are left the same size but simply moved to separate locations, each section will still be diluted with low-interest-potential books and if the browser's immediate interests involve both mathematics and physics, for example, he will simply have to walk further to scan sections which are still of low interest potential. Somehow the reconstruction of the library should lead back to more sections of high interest potential for the average user.

The solution is not simple, but it must involve a certain amount of concentration of high-interest-potential books in some subject sections. Sections which have become too large to browse through efficiently in the time the average browser can spend, should be separated, not into subdivisions according to subject, but into a high-interest section and a low-interest section. In other words, some fraction of the books in this overlarge section should be "retired" to a less accessible region of the library. This may perhaps be a disadvantage to a few users, who may be interested in the older or more specialized books (though it may also be advantageous to him), but it will definitely be of advantage to the majority of the users, who can again browse efficiently.

A word needs to be said here about the size and coverage of the subject sections we speak of here. A few specialist users will want to scan only those shelves covering the history of the reign of Philip Augustus, for example, but the majority, if they go to the history shelves at all, would tend to scan all books on French history, or even all European history. That subject section which the average user, in one of his visits, rates as having uniform interest potential, is what we shall call a uniform subject section. To the average browser the book he might want may be anywhere in such a section and he will tend (if the section has not grown too large) to scan it uniformly if he scans it at all. Data on actual usage (and correlation of usage) might be collected to decide how wide a subject range should be included in a uniform subject section, for a particular library. But most librarians, as well as many habitual users of a library, can make estimates of appropriate subject range which would be the right order of magnitude.

Until further measurements are made, we might assume that broad subject categories (such as physics or economics or ancient history) would correspond to uniform subject sections.

Returning again to the main problem, we reiterate that whenever a uniform subject section becomes too large for the average browser to cover effectively in a quarter to a half hour (larger than about 1000 to 2000 volumes) it should be split into a low-use section of "retired" books and a high-use section for browsing. It is not difficult to measure the degree of use of any individual book; if circulation is allowed, a book's circulation rate is a fairly good measure of its "popularity". Thus it is reasonable to consider that the average interest potential, for the average browser, for a given uniform subject section, is proportional to the mean circulation rate of the books in the section. If the librarian can make his split so as to have the mean circulation rate of the browsing portion considerably higher than that for the less accessible part, he will have made the task of the average browser much more rewarding. As mentioned before, optimal allocation of search effort is highly non-linear; a split which raises the interest potential by as little as 50 percent may make it worthwhile for many more browsers to scan it, though they would (and should) have ignored the previous, unseparated section.

It is thus assumed that the mean circulation $\bar{R}$ of the books in a section is proportional to the mean value of the interest potential of the section for the users who scan it at all;

$$\bar{V} = C\bar{R} \qquad \text{or} \qquad \bar{E} = C(N\bar{R}) \qquad (6)$$

Here $\bar{E}$ is the mean value of the a priori estimate of books in the section that are likely to be of immediate interest to the browser, averaged over those who browse in the section; $\bar{V} = \bar{E}/N$ is the mean interest potential, averaged over the same users; and $N\bar{R}$ is the total yearly circulation of the section. The exact value of constant C is not important for our present uses; we can conveniently take it to be about 0.001. We also assume that the chance of a particular book being the one a browser picks out is similarly proportional to the particular book's yearly circulation, R;

$$v = CR \quad ; \quad C \simeq 0.001 \qquad (7)$$

where v might be called the book potential of the individual book. Thus $\bar{V}$ is the average of the individual book potentials of all the books in the section.

There has been some study of the distribution of books according to their circulation[5]. For the purposes of this paper we need not go into detail, since changes of factors of 1.5 or 2 are the only ones worth considering here. To this approximation we can assume that the number of books, in a uniform subject section, which have book potential greater than v, is $Ne^{-v/\bar{V}}$, where N is the number of books in the section and $\bar{V}$ is given by Eq.(6). Thus the estimated number of books in the section with book potential between v and v + dv is

$$Nf(v)dv = \frac{N}{\bar{V}}\, e^{-v/\bar{V}}dv \quad ; \quad \int_0^\infty f(v)dv = \frac{1}{\bar{V}}\int_0^\infty e^{-v/\bar{V}}dv = 1 \qquad (8)$$

where f(v) is the probability density that a book has book potential in dv at v. The mean interest potential of the section is thus

$$\int_0^\infty v\, f(v)\, dv = \int_0^\infty \frac{v}{\bar{V}}\, e^{-v/\bar{V}}dv = \bar{V} = C\bar{R} \qquad (9)$$

An Optimal Retirement Plan

Thus the most effective way to separate off a high-interest-potential, browsing section from an overlarge subject section would be according to circulation rate. The "retired" section would be the fraction $(1-x)$ of books having book potential ranging from 0 to v_0 and the "reconcentrated" section would be the remaining xN books (where N is the size of the undivided section), each one having book potential greater than v_0. Of course v_0 is related to x by the requirement that

$$\int_{v_0}^{\infty} f(v)\,dv = e^{-v_0/\bar{V}} = x \quad \text{or} \quad v_0 = \bar{V}\ln(1/x) \qquad (10)$$

The mean interest potential of the "reconcentrated" section is then

$$\bar{V}_r = \frac{1}{x}\int_{v_0}^{\infty} v\,f(v)dv = \bar{V} + v_0 = \bar{V}\left[1 + \ln(1/x)\right] \qquad (11)$$

indicating an enhancement of the mean interest potential by the factor in square brackets. The mean interest potential of the "retired" section is

$$\bar{V}\left[1 - \frac{x}{1-x}\ln\left(\frac{1}{x}\right)\right]$$

displaying a corresponding reduction (not very much if x is small).

The mean interest potential of the "reconcentrated" section is increased, but at the cost of reducing the total number of books to be scanned and inevitably of reducing the estimated number of books of immediate interest to the average browser, for $\bar{E}$ equals $\bar{V}$ times the number of books in the section.

$$\bar{E}_r = xN\bar{V}_r = \bar{E}x\left[1 + \ln(1/x)\right] \qquad (12)$$

where $\bar{E} = N\bar{V} = C(N\bar{R})$ is the estimated number of books of interest[6] to the average browser in the original, undivided section. The quantity $x[1 + \ln(1/x)]$ is less than unity for all x between 0 and 1. Even if we retire only books with the lowest circulation rate, we will always retire some books which, once in a while, would be of interest to some browser. The reduction is not very great if we retire only a few books (i.e., if x is nearly unity) but then we would not have increased the mean interest potential by very much. The reduction becomes quite apparent if x is quite small; we would have increased the mean interest potential of the remaining books at the expense of depriving the browser of the chance to see a number of books he might occasionally be interested in.

To find the optimal middle ground we have recourse again to the search formulas (3) and (4), which hold for each browser. If the reconcentrated section is still so large that the average browser cannot efficiently scan the whole section in the time he can spend, then the fact that $\bar{E}_r$ is larger will not help, for he hasn't the time to find the books of interest among all the others. If x is made too small the average browser will "oversearch" the small collection, but will miss some of the books which have been retired. Somewhere between is an optimum size that will maximize the expected success for the average browser.

If the browser spends time t in the reconcentrated section his expected success is

$$S_r = \bar{E}_r(1 - e^{-\rho t/xN}) = \bar{E}x\left[1 + \ln(\tfrac{1}{x})\right](1 - e^{-\rho t/xN}) \qquad (13)$$

where ρ is the average of the browser's search rate and xN is the size of the reconcentrated section. Data on length of stay in the library indicates that it is distributed exponentially[5]. If the mean time spent browsing in the section under study is $\bar{T}$, the probability that a person spends between t and t + dt there during one visit is $(1/\overline{T})e^{-t/\overline{T}}dt$ and the mean value of the search factor $(1 - e^{-\rho t/xN})$ is

$$\frac{1}{\overline{T}}\int_0^\infty (1 - e^{-\rho t/xN})e^{-t/\overline{T}}dt = 1 - \frac{1}{1 + (\rho\overline{T}/xN)} = \frac{(\rho\overline{T}/N)}{x + (\rho\overline{T}/N)}$$

Thus the mean value of the expected success, averaged over all browsers, for a reconcentrated section containing the fraction x, of the books in the original section, which have the higher circulation rate, is

$$\overline{S}_r = \overline{E}x\left[1 + \ln(\tfrac{1}{x})\right]\frac{\gamma}{x + \gamma} \quad ; \quad \gamma = \frac{\rho\overline{T}}{N} \qquad (14)$$

According to the earlier discussion, $\rho\overline{T}$ is the number of books which can be scanned with about 70 percent efficiency in time $\overline{T}$, the mean time a browser spends. And $\gamma = \rho\overline{T}/N$ is the fraction of the original, undivided section which the average browser can scan adequately (i.e., 70 percent) in the average time he allocates to this section. Parameter γ can, of course, be larger than unity, in which case the section is small enough so there is no need to subdivide it.

Figure 4 shows the behavior of the function $\overline{S}_r/\overline{E}$, for different values of γ. The search factor $\gamma/(x + \gamma)$, responding to the fact that the larger the section the less meticulous can be the search, is unity at $x = 0$ and decreases as x approaches 1, first slowly and then more rapidly. The factor $x[1 + \ln(1/x)]$, measuring the expected number of books of potential interest in the concentrated fraction x, rises quickly from zero as x rises from zero and approaches 1 as x approaches 1. The product, $\overline{S}_r/\overline{E}$, has a maximum somewhere between 0 and 1, unless γ, the search density, is very large (in which case the optimum value of x is unity and there is no need to divide the section). But if γ is less than about 2, the average browser, during his

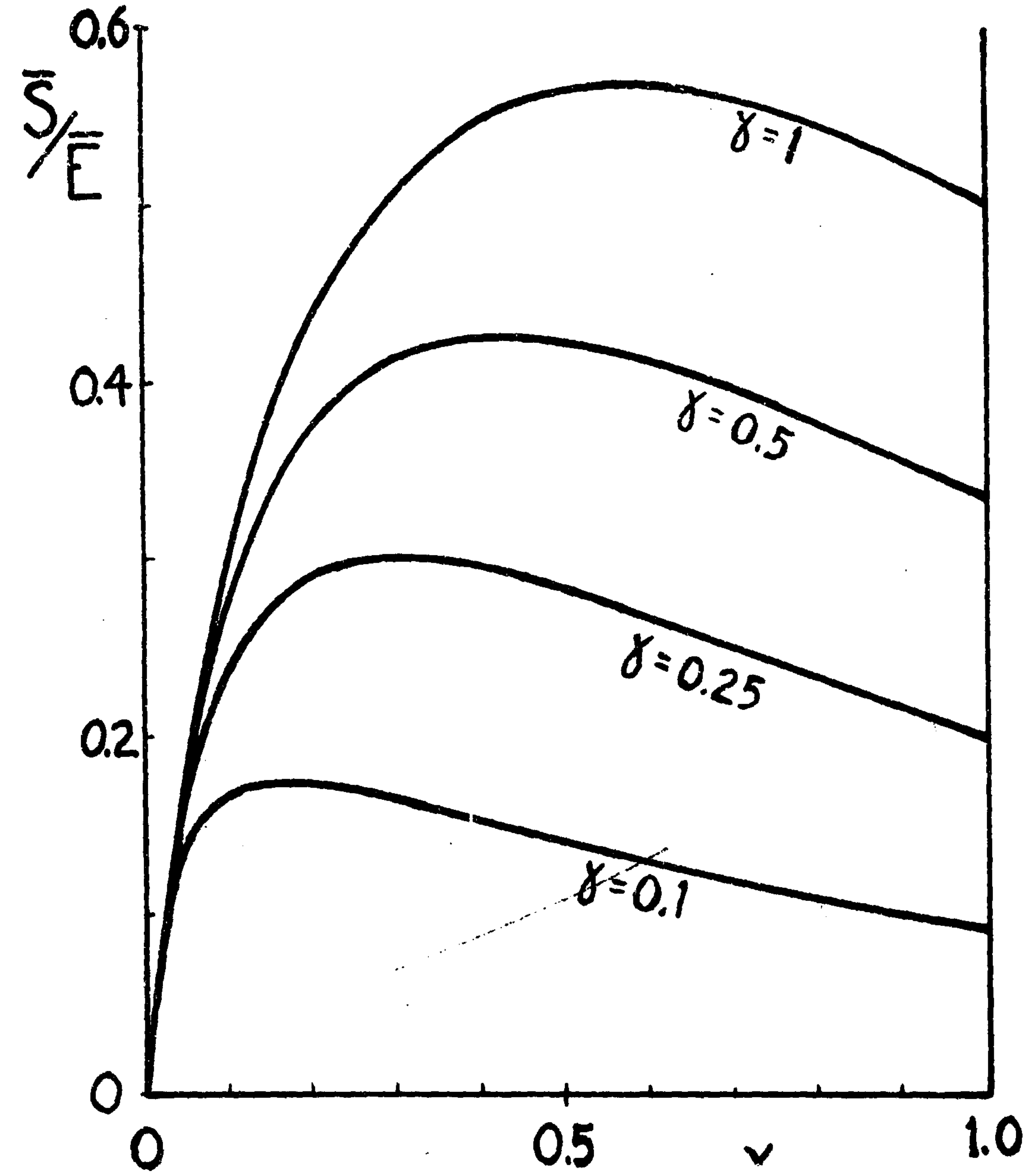

Fig.4. Changes in expected success as one changes the fraction x, of a uniform potential section, that is left in a reconcentrated, browsing section, when retirement criterion is low circulation.

average stay, cannot scan the full section effectively, the optimal value of x, x_o, is less than 1 and there is some advantage in breaking the section into a "retired" section and a "reconcentrated" section containing x_oN books. The advantage is not very great if γ is not much less than 1, but if γ is less than 0.1 the possible improvement is a factor of 2 or better, which is definitely advantageous for the browser.

Expression (14) can be differentiated to find, for different values of γ, the optimum value of x, the fraction of the original section which would yield the greatest success for the average browser. It is the solution of the equation

$$\frac{\ln(1/x)}{x+\gamma} = x\,\frac{1+\ln(1/x)}{(x+\gamma)^2} \quad \text{or} \quad x_o = \gamma \ln(1/x_o)$$

$$\text{and for this } x_o, \quad (\bar{S}_r/\bar{S}) = x_o\left(1+\frac{1}{\gamma}\right) \tag{15}$$

The optimal value of $\bar{S}_r$ is then $\bar{E}x_o$, which is to be compared with the value $\bar{S} = \bar{E}\gamma/(1+\gamma)$ for the undivided section ($x = 1$). This optimal browsing fraction x_o is plotted in Fig.5 and the advantage $\bar{S}_r/\bar{S}$ gained by the division is plotted in Fig.6, both as functions of γ.

A few examples may show how it can be used. Suppose ρ is 150 and $\bar{T}$ is 5 minutes (this may seem short, but many browsers scan several different sections in a visit) or $\rho = 100$ and $\bar{T}$ is 7.5 minutes; in any case $\rho\bar{T}$ is 750 and $\gamma = 750/N$. Now suppose the undivided section has $N = 1000$ volumes. In this case $\gamma = 0.75$ and the optimal x_o is about 0.5; we should retire half the collection. However we would only gain about 20 percent advantage ($\bar{S}_r/\bar{S} = 1.2$); it is doubtful whether this is worth the trouble of dividing the section. If there are about 2 books

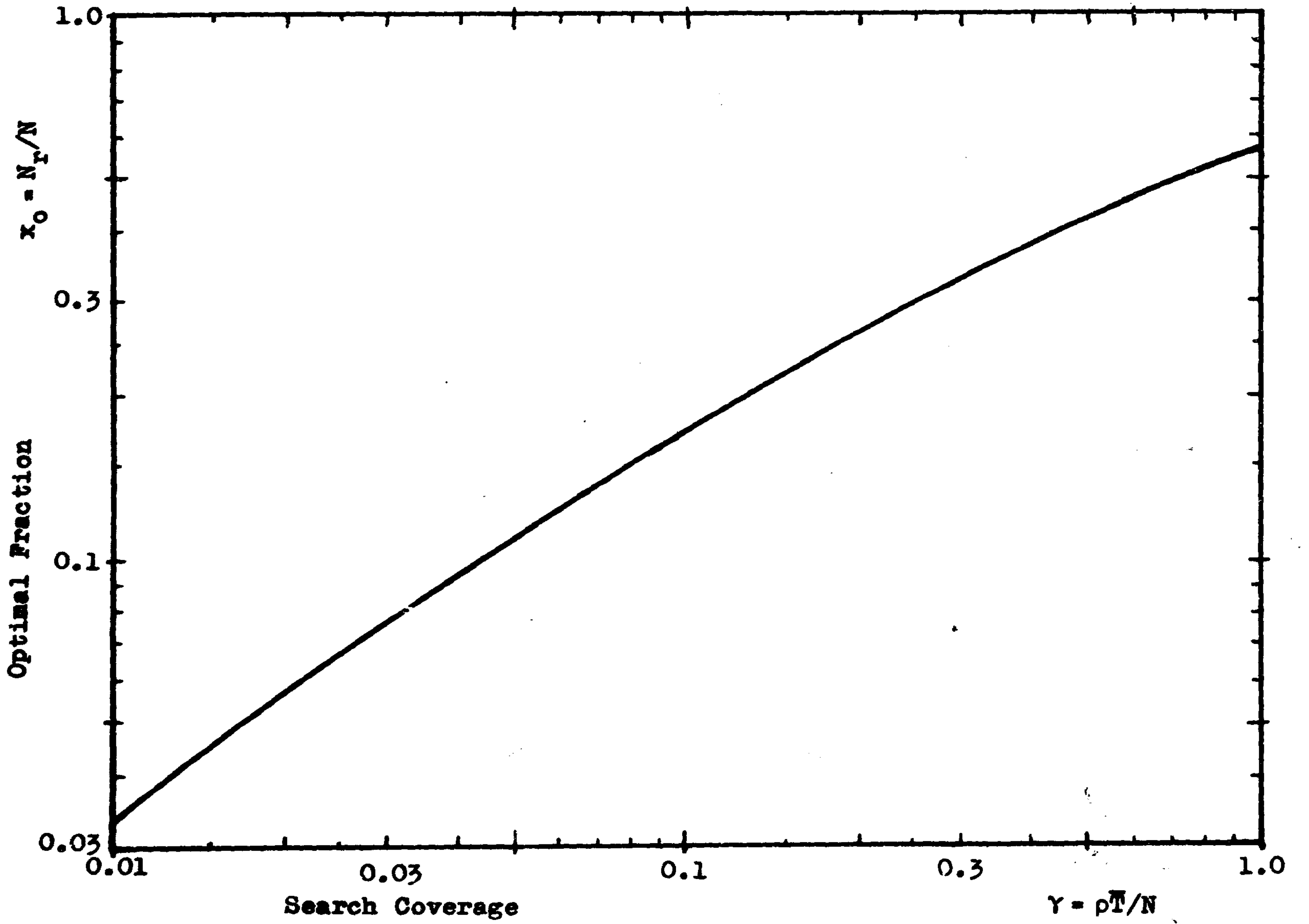

Fig.5. Optimal browsing fraction x_o as function of coverage parameter γ, for low-circulation retirement plan.

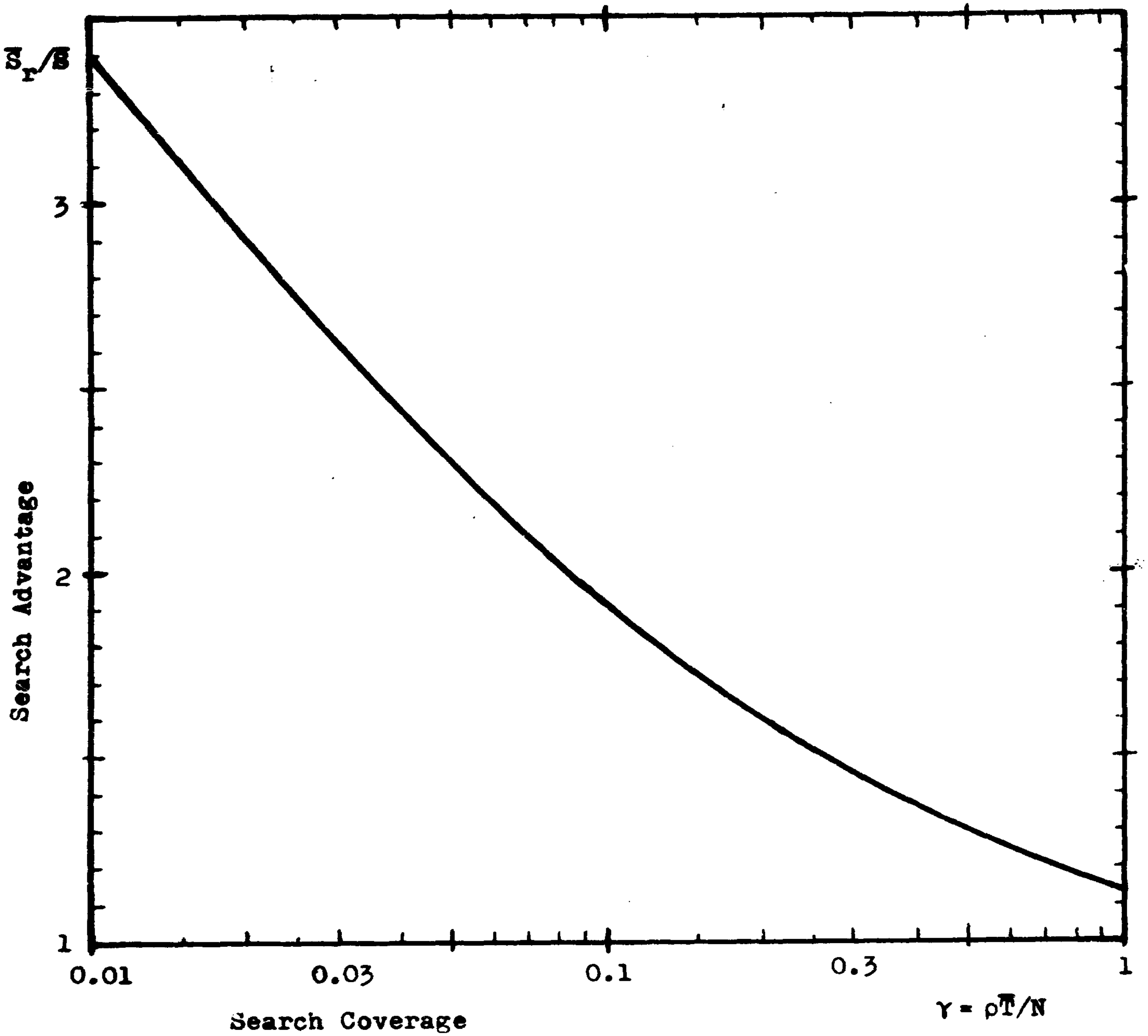

Fig.6. Search advantage as function of search coverage parameter γ, for the case of Fig.5.

of immediate interest per thousand volumes of the initial collection, the expected success would be about 1 volume either from the full section or the concentrated half-section.

On the other hand, if N were 5000, γ would be 0.15, for which $x_o = 0.25$; the optimal browsing section would be the most popular 1200 books out of the 5000. The change in success expected would be by a factor $\bar{S}_r/\bar{S} = 1.7$, probably a worthwhile improvement. If there were about 1 book of interest per 500 volumes,for the average browser,in the original section, $\bar{S}$ for the full section would be about 1.5; $\bar{S}_r$ for the quarter-sized, reconcentrated section would be 2.5.

Finally, suppose N were 30,000, all in a collection homogeneous enough to make it difficult for the average user to distinguish one part from another for browsing. The formulas indicate that this outsize section should be thinned by retirement to a browsing collection of the 2000 highest-circulation books, which would have a search advantage over the full collection of a factor of 2.7, raising the expected average success from $\bar{S} = 1.5$ for the full collection to $\bar{S}_r = 4$ for the reconcentrated section, a change definitely advantageous for the average browser. As mentioned earlier, a search advantage of 1.5 or better (corresponding to a γ of 0.25 or smaller) would probably justify dividing a subject section, <u>if</u> browsing is an important factor <u>and</u> <u>if</u> the division can be made according to circulation.

Retiring Books by Age

The procedure of retirement according to circulation is probably optimal. However continual retirement of low-circulation books from a browsing section demands a greater awareness of book circulation than most libraries have at present. Let us see what can be done if the over-size section is divided on the basis of age. Measurement[5] has shown that the distribution in circulation (and thus in potential interest) of books which have been on the shelf for t years is, very approximately,

Probability that a book of shelf age t has a book potential greater than v is $e^{-v(t+t_o)/v_o t_o}$

where v_o is the book's potential during its first year on the shelf and t_o is a parameter typical of the class of book and of the average user of the library. In a science library t_o may be 1 for physics books and 2 for mathematics books, for example; books on history may have t_o as large as 10 or 20. The mean book potential for a book of shelf age t would be $v_o t_o/(t+t_o)$ if v_o were its potential during its first shelf year and t_o were the parameter for books of its class. Thus the larger t_o is, the slower does the class of books decrease in book potential.

If the uniform-interest section under study contains books more or less equally distributed in age from the most recent acquisitions to the oldest with shelf age t_m, then the probability that a book, taken at random from the section, has book potential v or greater is

$$P(\geqslant v) = \frac{1}{t_m}\int_0^{t_m} e^{-v(t+t_0)/\bar{V}_0 t_0}\,dt \qquad (16)$$
$$= \frac{\bar{V}_0 t_0}{v t_m}\left[e^{-(v/\bar{V}_0)} - e^{-(v/\bar{V}_0 t_0)(t_0+t_m)}\right]$$

where $\bar{V}_0$ is the mean book potential of all books of the section during their first year of shelf life. The probability density corresponding to the $f(v)$ of Eq.(11) is $f(v) = -dP(\geqslant v)/dv$.

The mean interest potential of this collection of books is

$$\bar{V} = \int_0^\infty v\,f(v)dv = \int_0^\infty P(\geqslant v)dv = \frac{\bar{V}_0 t_0}{t_m}\ln(1 + \frac{t_m}{t_0}) \qquad (17)$$

which is less than $\bar{V}_0$ if t_m is larger than t_0, i.e., if there is an appreciable fraction of older books in the section. Now suppose we pick from these a browsing collection, with xN volumes, by keeping all the books of shelf age xt_m or less and retiring the rest to a less accessible location. It is not difficult to see that the mean interest potential of this collection is

$$\bar{V}_r = \frac{\bar{V}_0 t_0}{x t_m}\ln(1 + x\frac{t_m}{t_0}) = \frac{\bar{V}}{x}\,\frac{\ln[1 + x(t_m/t_0)]}{\ln[1 + (t_m/t_0)]} \qquad (18)$$

and the expected number of books of immediate interest to the average browser is $\bar{E}_r = xN\bar{V}_r$. Substituting, we see that

$$\bar{E}_r = \bar{E}\,\frac{\ln[1 + x(t_m/t_0)]}{\ln[1 + (t_m/t_0)]} \qquad (19)$$

where $\bar{E} = \bar{V}N$ is the expected number of books of immediate interest in the full collection. This drops in value from $\bar{E}$ to zero as x goes from 1 to 0, just as does the factor $x[1 + \ln(1/x)]$ of Eqs.(12) and (14).

Again we introduce the search factor, as in Eq.(14), and obtain the expected success if the average browser spends average time $\bar{T}$ scanning the reconcentrated section,

$$\bar{S}_r = \frac{E\gamma}{\ln(1+\beta)}\,\frac{\ln(1+x\beta)}{x+\gamma}\ ;\quad \beta = \frac{t_m}{t_o}\ ;\ \gamma = \frac{\rho\bar{T}}{N} \qquad (20)$$

As with the function of Eq.(14), for the "retirement by circulation procedure", this function has a maximum, at $x = 1$ if γ is somewhat larger than 1, for x less than 1 if γ is less than 1. The equations,giving the optimal value x_o and the expected success for the full collection, $\bar{S}$, and for the reconcentrated section, $\bar{S}_r$, are

$$\beta\gamma = (1 + x_o\beta)\ln(1 + x_o\beta) - x_o\beta \longrightarrow \tfrac{1}{2}(\beta x_o)^2 \quad (\beta\gamma < 0.1)$$
$$\bar{S} = E\,\frac{\gamma}{1+\gamma}\ ;\quad \bar{S}_r = \bar{S}\,\frac{\beta(1+\gamma)}{(1+x_o\beta)\ln(1+\beta)} \qquad (21)$$

Values of βx_o are plotted against $\beta\gamma$ in Fig.7 for this less efficient retirement plan, and values of the expected search advantage $\bar{S}_r/\bar{S}$ are shown in Fig.8. Again a few examples are in order, in order to compare results with those of Eq.(15) for the more efficient plan. Here we must distinguish between rapidly aging books (t_o small, β large) and slowly aging books (t_o large, β small). We assume , for the example, that the undivided section has books of all shelf age from 0 to 20, so $t_m = 20$. We take $\beta = 2$ ($t_o = 10$) for the slowly aging example and $\beta = 10$ ($t_o = 2$) for the rapidly aging example and, as before, we assume that $\rho\bar{T} = 750$, so that $\gamma = 750/N$. For comparison we assume that there are about 2 volumes,of immediate interest to the average browser, per 1000 volumes in the undivided section.

The section with 1000 books will certainly not be worth dividing by age, so we start with the example where the undivided section has 5000 books, so $\gamma = 0.15$, $E = 10$ and $\bar{S} = 1.3$. For the case $\beta = 2$, $\beta\gamma = 0.3$, $\beta x_o = 0.9$, $x_o = 0.45$ and $\bar{S}_r/\bar{S} = 1.1$;

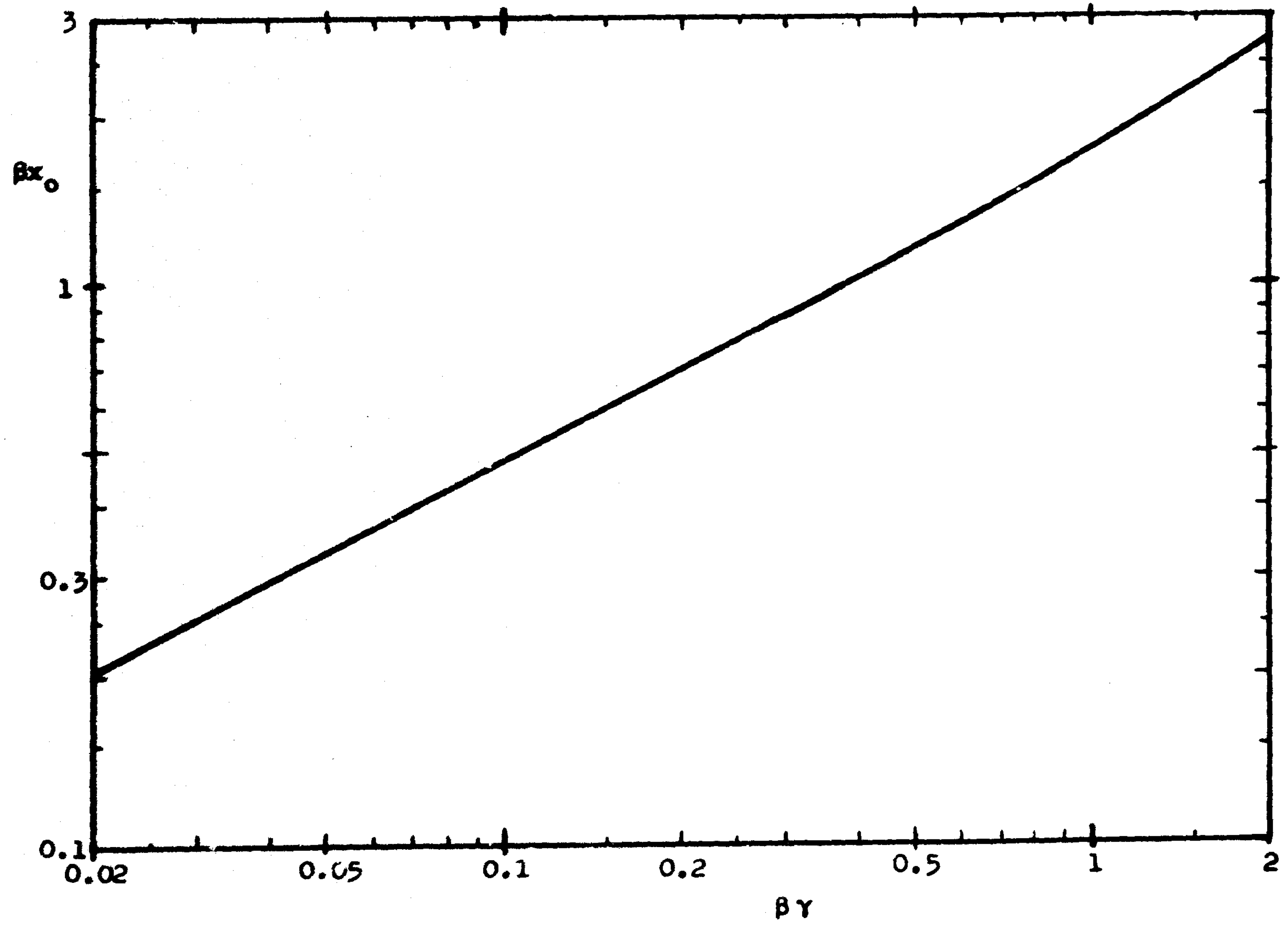

Fig.7. Optimal fraction x as function of coverage parameter γ and aging parameter β, when retirement criterion is shelf life of book.

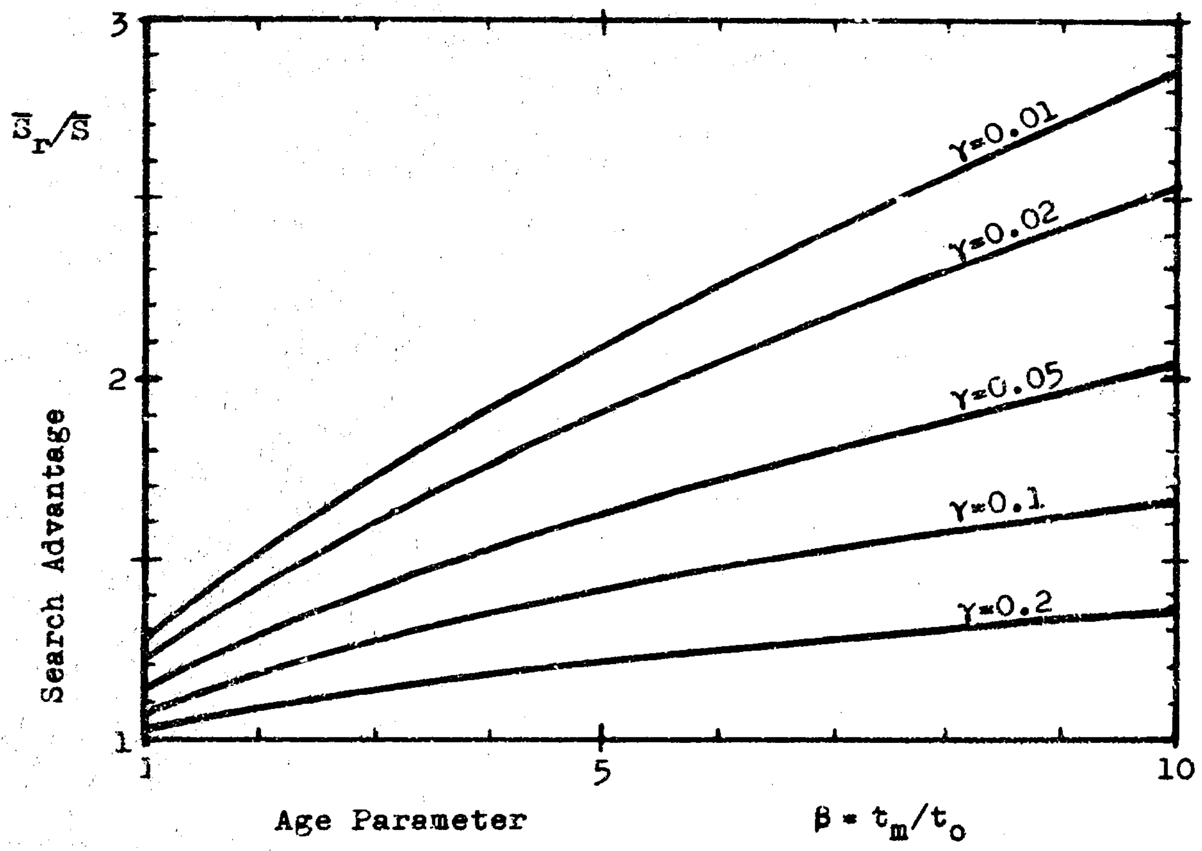

Fig.8. Search advantage for different values of coverage γ, as function of age parameter β, for case of Fig.7.

the gain from $\bar{S} = 1.3$ to $\bar{S}_r = 1.4$ by retiring the oldest half of the collection, is not worth the trouble. On the other hand if the collection was rapidly aging, $\beta = 10$, then $\beta\gamma = 1.5$, $\beta x_o = 2.5$, $x_o = 0.25$ and $\bar{S}_r/\bar{S} = 1.4$; the gain from 1.3 to 1.8 in S may be worth the trouble of retiring all but 1250 of the youngest books. Comparison with the similar examples in the earlier discussion shows that retirement by age is not as good for browsers as retirement by circulation, particularly for the slowly-aging books (β less than 5).

Even if the uniform subject section consists initially of 30,000 volumes, separation by age does not help much for the slowly-aging classes. Here $\gamma = 0.025$, $\bar{E} = 60$ and $\bar{S} = 1.5$ as before. For $\beta = 2$ we have $\beta\gamma = 0.05$, $\beta x_o = 0.34$, $x_o = 0.17$ and $\bar{S}_r/\bar{S} = 1.4$; increasing expected success from 1.5 to 2 by retiring all but 5000 of the youngest volumes may just be worth while (particularly if the collection must be split because of shortage of space). On the other hand if the collection is rapidly aging, with $\beta = 10$, then $\beta\gamma = 0.25$, $\beta x_o = 0.8$, $x_o = 0.08$ and $\bar{S}_r/\bar{S} = 2.4$; reducing the browsing collection to the roughly 2500 books no more than 2 years old will increase the expected success for the average browser from 1.5 to 3.6, a factor of 2.4. This is a definite gain, though not as great as the factor 2.8 which would be obtained if the separation were on the basis of circulation. Retirement by age retires some high-potential books simply because they are older and leaves in the browsing collection too many low-potential books simply because they are younger. However the efficiency of separation by age is not too bad for rapidly-aging book classes. If we say that a search advantage $\bar{S}_r/\bar{S}$ of 1.5 or more makes it worth while to separate off a browsing section then, for classes with $\beta = 10$, γ should be less than 0.15, for $\beta = 2, \gamma$ should be less than 0.01 before browsing advantage would make separation worth while.

These criteria will become easier to determine the more accurately the parameters ρ, β, $\bar{T}$ are determined for the library and its users, and the more precisely one can delimit the various uniform subject sections of the library. Also the separation of the over-large sections into high-potential, browsing sections and more-secluded sections for the less-used volumes, will become easier when means are devised, by computer or otherwise, to keep continuous track of the circulation rates of the books in the library.

References and Notes

[1] For the application of search theory to antisubmarine search, see P.M.Morse and G.E.Kimball, *Methods of Operations Research* , MIT Press and John Wiley and Sons, New York, 1951, and B.O.Koopman, Operations Research, **4**,296,1956 and **4**,503,1956.

[2] This formula, in the context of antisubmarine search, is derived in B.O.Koopman, Operations Research,**5**,613,1957, which summarizes work carried out during World War II.

[3] See, for example, H.Chernoff and L.Moses, *Elementary Decision Theory*, John Wiley and Sons, New York, 1959.

[4] For further discussion of this point, see the paper "Search and Browsing" by P.M.Morse in the forthcoming "Festschrift" for Jesse Shera, to be published by the Case Western Reserve Press.

[5] See, for example, P.M.Morse, *Library Effectiveness* , MIT Press, Cambridge, Mass., 1968 and other references given there.

[6] Of course the individual books which would be of immediate interest to one browser would differ from those of immediate interest to another. If neither browser knows where in the section his "wanted" volumes are, we are justified in taking the average of the number of such books in the section, as estimated by each browser (even though each estimate refers to different books),as the value of $\bar{E}$ for the section.

Unclassified
Security Classification

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1 ORIGINATING ACTIVITY *(Corporate author)*
M.I.T. Operations Research Center
77 Massachusetts Avenue
Cambridge, Massachusetts 02139

2a. REPORT SECURITY CLASSIFICATION
Unclassified

2b. GROUP

3 REPORT TITLE

ON BROWSING: THE USE OF SEARCH THEORY IN THE SEARCH FOR INFORMATION

4 DESCRIPTIVE NOTES *(Type of report and inclusive dates)*
Technical Report No. 50, February 1970

5 AUTHOR(S) *(First name, middle initial, last name)*

Morse, Philip M.

6 REPORT DATE	7a. TOTAL NO. OF PAGES	7b. NO. OF REFS
February 1970	37	6

8a. CONTRACT OR GRANT NO
DA-31-124-ARO-D-209

b. PROJECT NO
20011501B704

c.

d.

9a. ORIGINATOR'S REPORT NUMBER(S)
Technical Report No. 50

9b. OTHER REPORT NO(S) *(Any other numbers that may be assigned this report)*
DSR 75217

10 DISTRIBUTION STATEMENT

Releasable without limitations on dissemination.

11 SUPPLEMENTARY NOTES
U.S.Army Research Office-Durham
Box CM, Duke Station
Durham, North Carolina 27706

12 SPONSORING MILITARY ACTIVITY

13 ABSTRACT

Search theory, originally developed for antisubmarine search, is applied to the scanning of library shelves for books of interest, or of a computerized abstract catalogue for items of immediate application. Procedures for optimizing the information to be found are discussed, as well as methods whereby the operational parameters can be measured. The organization and reorganization of a library, or other informational system, so as to improve its response to a searcher, are treated and curves are provided which indicate the degree and nature of the reorganization which can optimize this improvement. (U)

Unclassified
Security Classification

14. KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Library						
Information						
Search theory						
Computerized library catalogues						